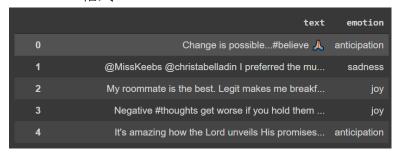
106020025

張瀷鏵

I. Data preprocessing

- A. 輸入資料
 - 1. 讀取一個 json 和兩個 csv 檔案
 - 2. 建立兩個 dictionary
 - a. {Id:emotion}
 - b. {id:identify}
 - 3. 將 json 的 id 掃過一遍
 - a. If b[id] == 'train' => add text and emotion
 - 4. Dataframe 格式



II. Feature engineering

A. tfidf 進行 vectorize(tokenizer 用 nltk, max featrue 取 1000)

III. Model

A. Deep learning:

跟 HW2 的 Deep learning 差不多,只是我將 tfidf 的 max_feature 增加到 1000 個,所以 input layer 為 1000,output layer 為 8,最後的結果為 0.416。

B. Bert:

沒有成功製作出來,嘗試了很多不同種的方式,最後都因為記憶體大小的限制導致無法正確的輸出模型,可能是因為這次的資料數量龐大,或是我對記憶體的控管不當所導致的。有嘗試過 sample 出資料的一小部分單獨做 Bert,但是準確度卻異常的低,原因除了資料量較小以外,也有可能是 tokenize 的問題。因為在使用 Bert 進行 training 時,所用的 tokenize 是 Bert 內建的,表情符號可能會被排除在外,因此損失了重要的 feature。